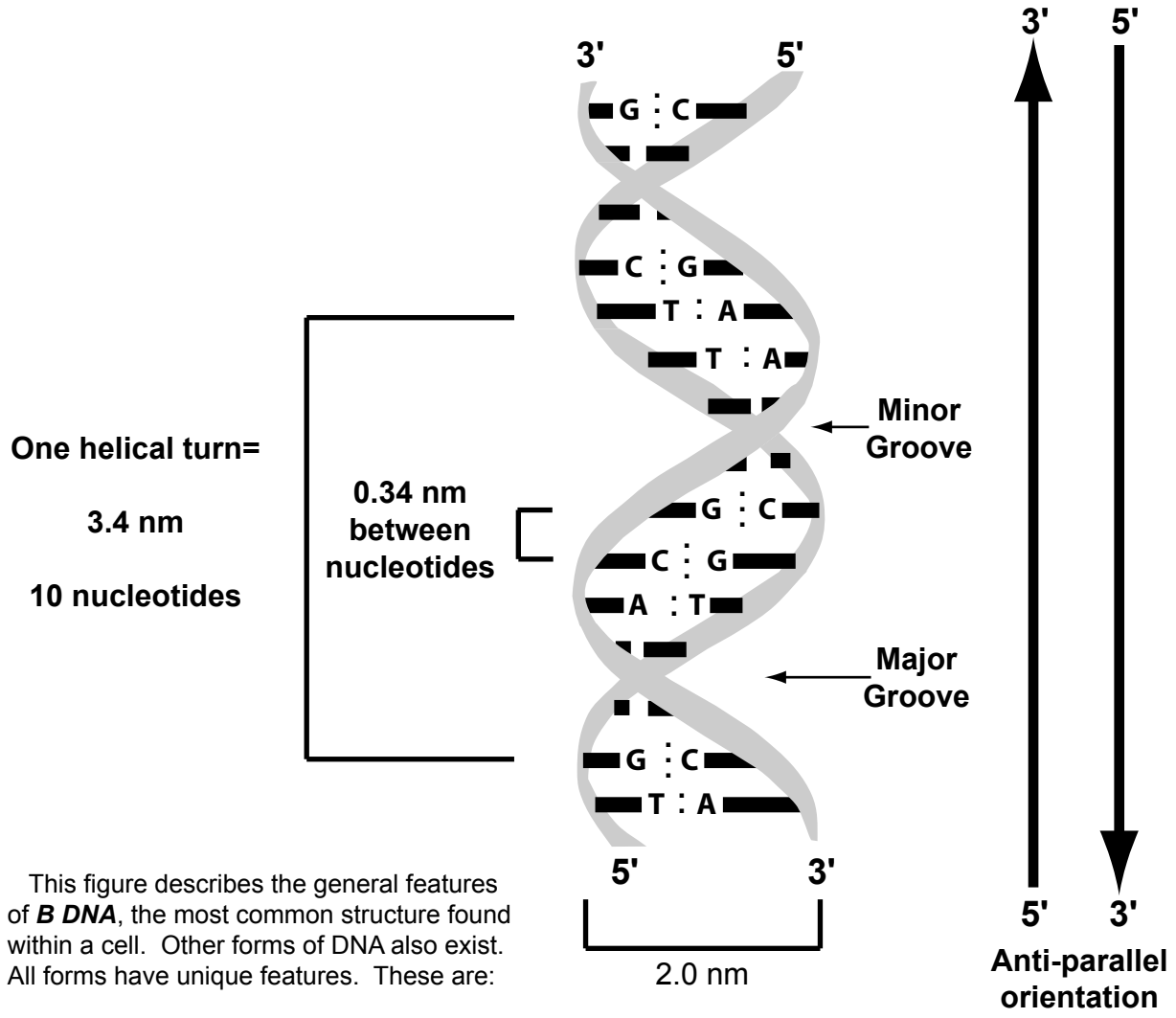
DNA Structure

A. The Concept

DNA has a regular structure. It's orientation, width, width between nucleotides, length and number of nucleotides per helical turn is constant. All of these features were described by Watson and Crick. Adenine is always opposite thymine, and cytosine is always oppostie guanine. The two strands are held together by hydrogen bonds: two bonds between adeninine and thymine and three bonds between guanine and cytosine.



This figure describes the general features of ***B DNA***, the most common structure found within a cell. Other forms of DNA also exist. All forms have unique features. These are:

Form	Helix Direction	Nucleotides per turn	Helix Diameter
A	Right	11	2.3 nm
B	Right	10	2.0 nm
Z	Left	12	1.8 nm

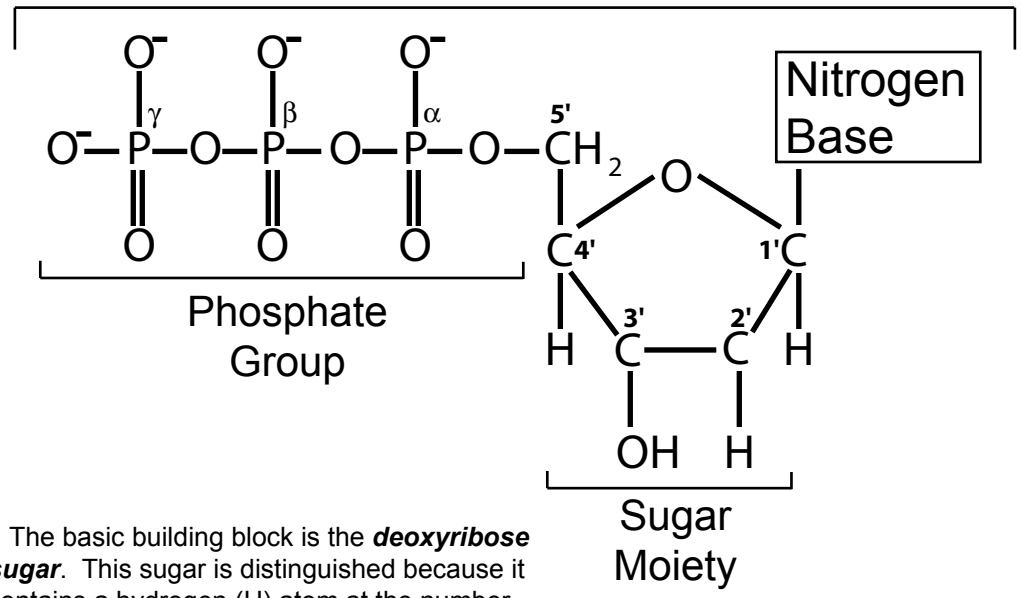
Figure 3. The structure of common DNA molecules.

Deoxyribonucleotide Structure

A. The Concept

DNA is a string of deoxyribonucleotides. These consist of three different components. These are the ***dexoyribose sugar***, a ***phosphate group***, and a ***nitrogen base***. Variation in the nitrogen base composition distingushes each of the four deoxyribonucleotides.

Basic deoxyribonucleotide components



The basic building block is the ***deoxyribose sugar***. This sugar is distinguished because it contains a hydrogen (H) atom at the number 2' carbon. Normal ribose has a hydorxyl (-OH) group at this position.

Attached to the 5' carbon is a triphosphate group. This group is important because in a DNA chain it undergoes a reaction with the 3' OH group to produce polydeoxynucleotide.

The final feature of the molecule is a ***nitrogen base***. These are attached to the 1' carbon. Four bases are possilbe. Two pyrimidines (thymie and cytosine) and two purines (adenine and guanine). The double stranded DNA molecule is held together by hyrodgen bonds. Pairing involves specific atoms in each base. Adenine pairs with the thymine, and guanine pairs with cytosine. These pairings and the atoms involved are shown to the right.

You have probaly heard of ATP, the energy moleucle. It is the deoxyribonucleotide to which adenine is attached. This molecule serves two very important functions in biological organisms.

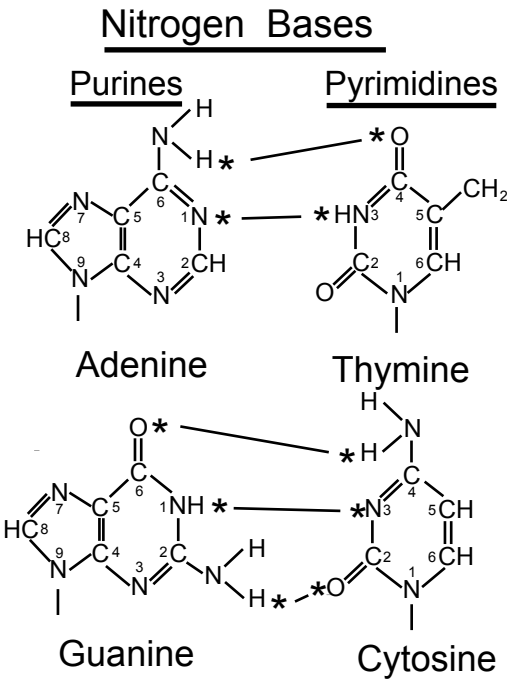
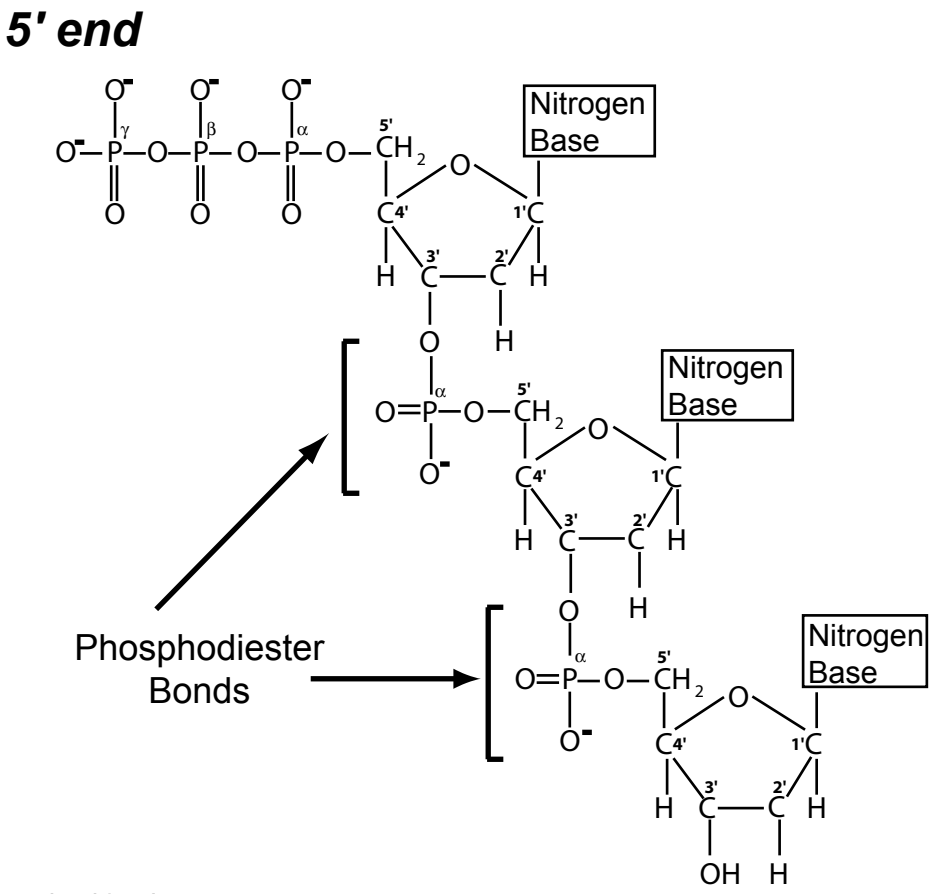


Figure 4. The structure of deoxyribonucleotides and base pairing among N bases.

A Single Strand Molecule of DNA

A. The Concept

Each strand of the double-stranded DNA molecule has the same basic structure. It is a series of series of deoxyribonucleotides linked together by phophodiester bonds.



DNA is a polynucleotide. It consists of a series of deoxyribonucleotides that are joined by phosphodiester bonds. This bond joins the a phosphate group to the 3' carbon of the deoxyribose sugar.

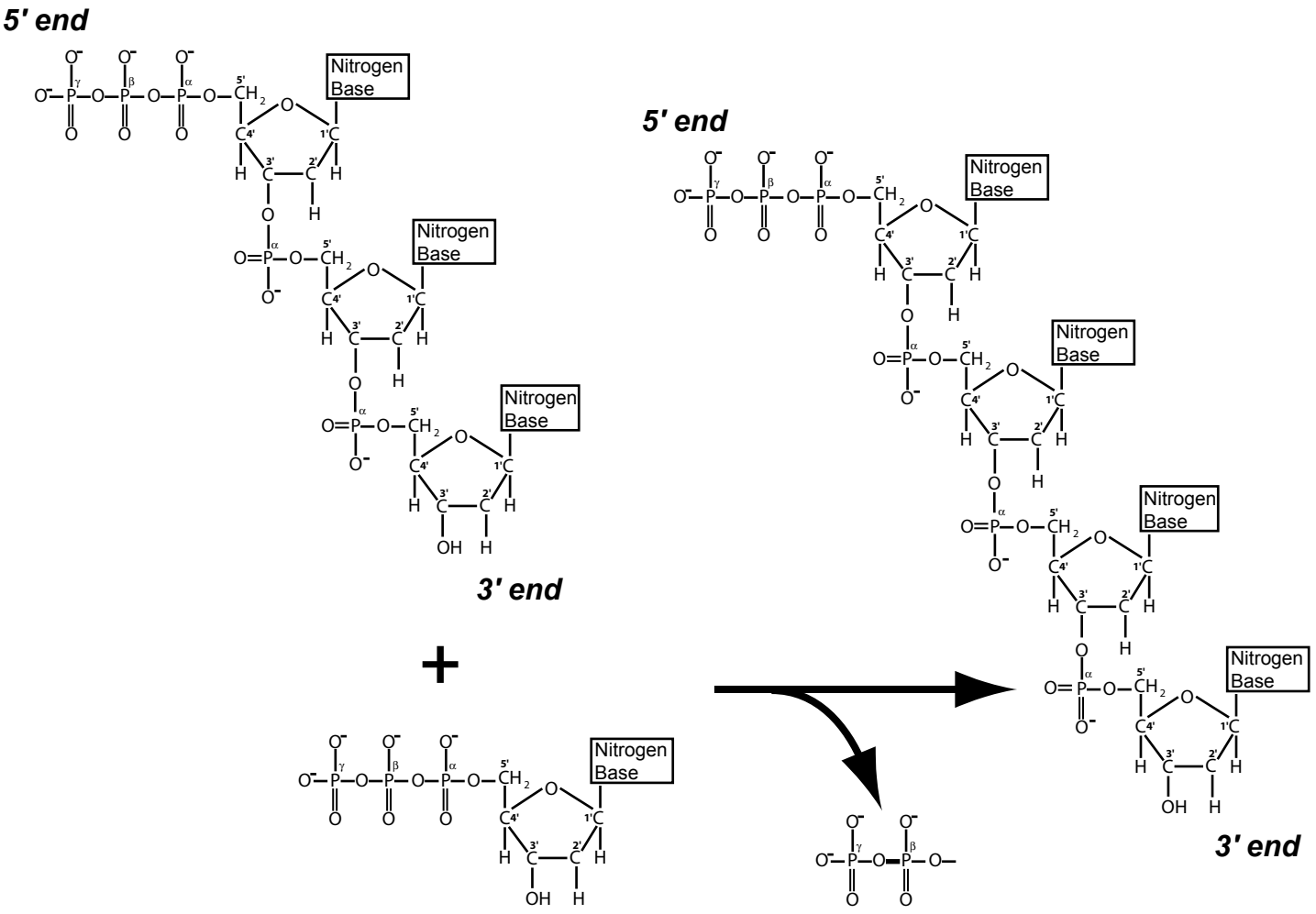
Each strand is complementary to the opposite strand. If one strand has an adenine at a position, its anti-parallele strand would have a thymine at the the corresponding position. Likewise, guanine and cytosine would be complementary.

Fig. 5. The single strand structure of DNA.

Making a Phosphodiester Bond/ Growing the DNA Chain

A. The Concept

The addition of a new nucleotide to a DNA molcule creates a phosphodiester bond. This requires the DNA chain that is being elongated and a deoxyribo-nucleotide.



Phosphodiester bonds are formed when a news dideoxynucleotide is added to a growing DNA molecule. During the reaction, a condensation reaction occurs between the α phosphate of the nucleotide and the hyroxyl group attached to the 3' carbon. This reaction is performed by the enzyme DNA polymerase. This is also an energy requiring reaction. The energy is provided by the breaking of the high-energy phophate bond in the nucleotide. This results in the release of a pyrophosphate molecule.

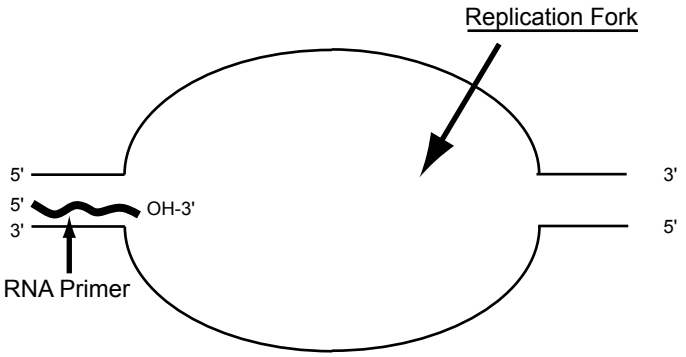
Figure 6. The formation of the phosphodiester bond that grows the DNA chain.

Steps of DNA Replication (Part 1)

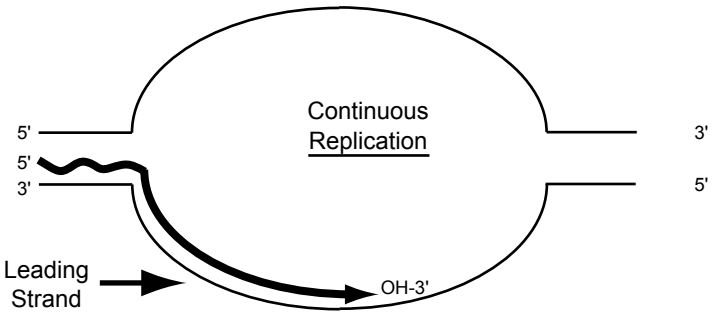
A. The Concept

DNA replication is essential biological process. It's primary function is to produce new DNA for cell division. The process has several distinct steps that are important to understand. The factors that are absolute requirements for DNA replication to begin are a ***free 3'-OH group*** and a ***DNA template***. A RNA primer provides the free 3'-OH group. The DNA to be replicated serves as the template. It is important to remember that ***all*** DNA replication proceeds in the 5'-3' direction.

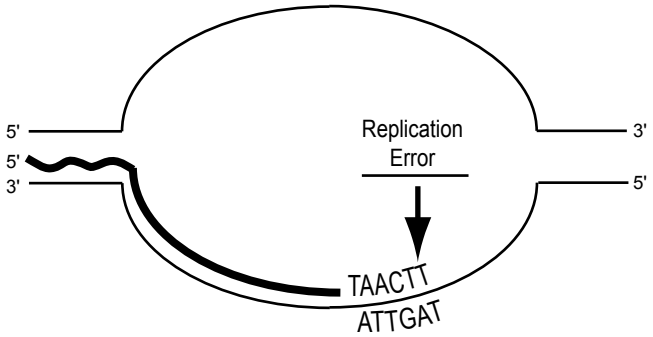
1. The replication fork is formed; RNA primer added.



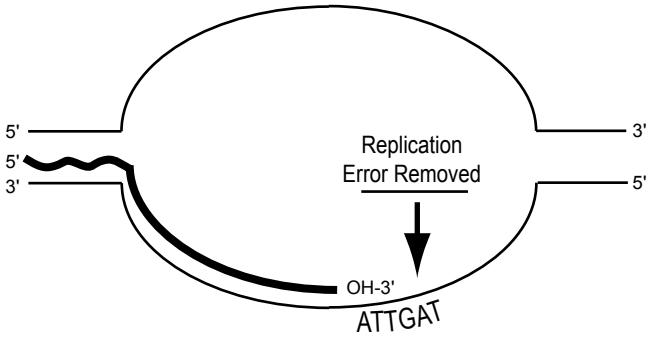
2. DNA is replicated by the 5'-3' synthesis function of DNA polymerase using the leading strand in a continuous manner.



3. An error occurs during DNA replication.



4. The DNA replication error is removed by 3'-5' exonuclease function of DNA polymerase.



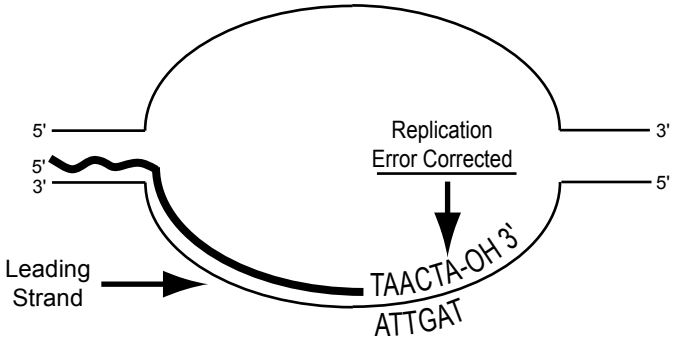
Notes on *E. coli* replication:

DNA Polymerase I and III. Pol III is the primary replicase enzyme that performs the elongation of the DNA strand. It adds nucleotides first to the RNA primer and then grows the chain by creating the phosphodiester bonds. It also has a 3'-5' proofreading (exonulcease) function that removes incorreclty incorporated nucleotides. DNA Pol I also has the 5'-3' replicase function, but it is primarily used to fill the gaps in the replicated DNA that occur when the RNA primer is removed. This enzyme also has a 5'-3' exonuclease function that is used to remove the RNA primer.

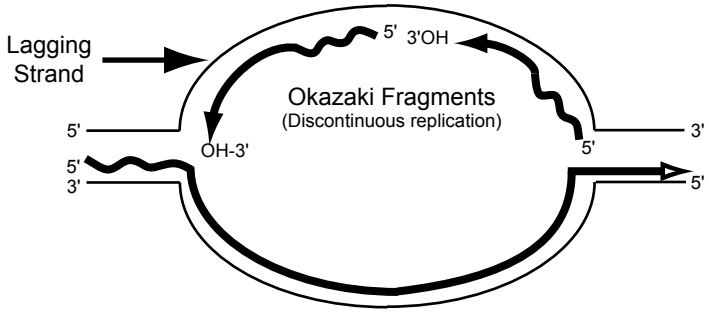
Figure 7. The steps of DNA replication.

Steps of DNA Replication
(Part 2)

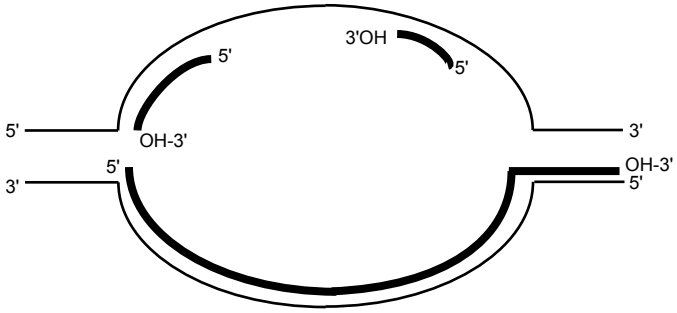
5. The DNA replication error is corrected.



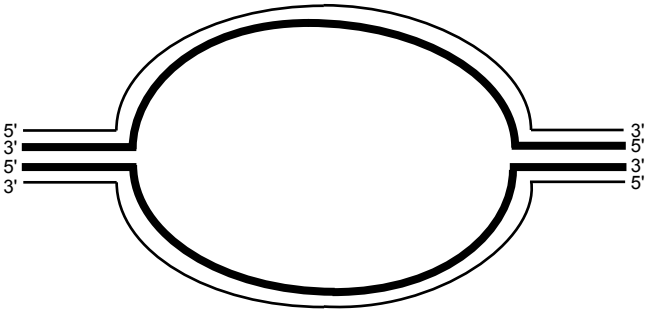
6. Meanwhile, Okazaki fragments are synthesized using the lagging strand in a discontinuous manner and leadng strand are completed simultaneously.



7. The RNA primers are removed by 5'-3 exonuclease function of DNA polymerase.



8.. Replication is completed by the filling in the gaps by DNA polymerae and DNA ligase.



Notes on replication:

Okazaki fragments: Both prokaryotic and eukaryotic DNA replication proceed in the 5'-3' direction. This poses a problem because the replication fork on moves in that direction. The problem relates to what is called the ***lagging strand***. It must be replicated in a direction that is opposite of the direction of the replication fork. This problem was solved by the discovery of Okazaki fragments (named after the person who discovered the process. In contrast to the ***leading strand***, in which DNA is replicated as a single molecule in a ***continuous*** manner, DNA is replicated in a ***disocontinuous*** manner on the lagging strand. Each of these is primer with a RNA primer, and DNA PolIII in E. coli makes short stretches of DNA. These fragments are then stitched together when the primer is removed and the strands completed by the action of DNA Pol I and ligase.

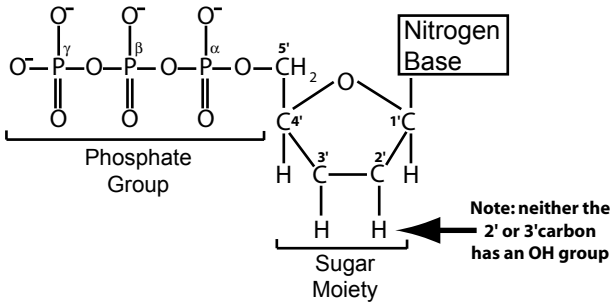
Figure 7 (cont.). The steps of DNA replication.

Chain Termination Sequencing: the Sanger Technique

A. The Concept

DNA sequencing is the most techique of genomics. By collecting the sequence of genes and genomes we begin to understand the raw material of phenotype development. The most common DNA sequencing is called ***chain termination sequencing*** or the ***Sanger technique*** (named after the person who created it). It is called chain termination because the incorporation of a ***dideoxynucleotide*** terminates the replication process because the nucleotide lacks the required 3'-OH group.

a. A dideoxynucleotide



b. The reaction reagents

DNA template
sequencing primer
dNTPs
ddNTPs (low concentration)
DNA polymerase
salts

c. The sequencing reaction result: fragments that differ by one nucleotide in length

Template → A T T C G G A T C C T T A A

Primer →

5' T A A G C C T A G G A A T T - H 3'
5' T A A G C C T A G G A A T - H 3'
5' T A A G C C T A G G A A - H 3'
5' T A A G C C T A G G A - H 3'
5' T A A G C C T A G G - H 3'
5' T A A G C C T A G - H 3'
5' T A A G C C T A - H 3'
5' T A A G C C T - H 3'
5' T A A G C C - H 3'
5' T A A G C - H 3'
5' T A A G - H 3'
5' T A A - H 3'
5' T A - H 3'
5' T - H 3'

When a dideoxynucleotide is inserted, the DNA replication process terminates because dideoxynucleotides do not have the necessary free 3' hydroxyl group required for the addition of additional nucleotides. This results in fragments that differ by one nucleotide in length.

Figure 8. The chain termination (Sanger) DNA sequencing technique.

Gel-based Detection of DNA Sequences

A. The concept

Four DNA sequencing reactions are performed. Each contains only one of the four dideoxynucleotides. Each reaction is added to a single lane on the gel. Since one of the dNTPs is radioactive, the gel in which the fragments are separated, can be used to expose an x-ray film and read the sequence.

a. The sequencing products

Reaction with ddATP

A T T C G G A T C C T T A A
5' T A A G C C T A G G A A -H 3'
5' T A A G C C T A G G A -H 3'
5' T A A G C C T A -H 3'
5' T A A -H 3'
5' T A -H 3'

Reaction with ddTTP

A T T C G G A T C C T T A A
5' T A A G C C T A G G A A T T -H 3'
5' T A A G C C T A G G A A T -H 3'
5' T A A G C C T -H 3'
5' T -H 3'

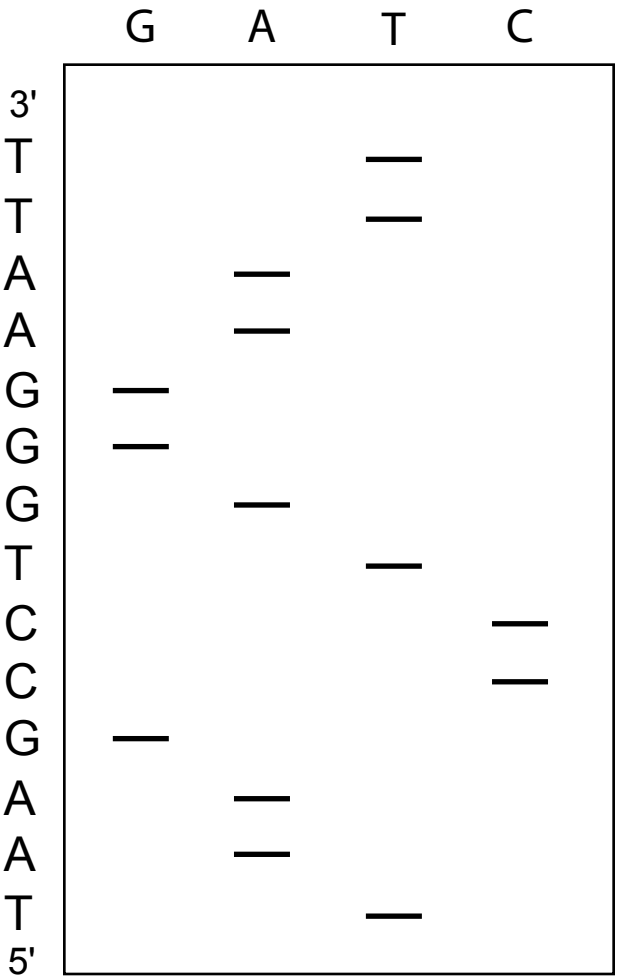
Reaction with ddGTP

A T T C G G A T C C T T A A
5' T A A G C C T A G G -H 3'
5' T A A G C C T A G -H 3'
5' T A A G -H 3'

Reaction with ddCTP

A T T C G G A T C C T T A A
5' T A A G C C -H 3'
5' T A A G C -H 3'

b. The sequencing gel



The sequencing reactions are separated on a polyacrylamide gel. This gel separates the fragments based on size. The shorter fragments run further, the longer fragments run a shorter distance. This allows the scientists to read the sequence in the 5'-3' direction going from the bottom to the top of the gel.

Figure 9. Gel-based detection of DNA sequencing products.

Fluorescent Sequencing and Laser Detection

A. The Concept

Rather than using four different reactions, each with a single dideoxynucleotide, the advent of fluorescently labeled dideoxynucleotide enabled 1) the sequencing reaction to be performed in a single tube, and the fragment could be detected by laser technology. Originally, the products were separated in a polyacrylamide gel prior to laser detection. The introduction of capillary electrophoresis, coupled with laser detection enabled the detection of up to 96 products at a time.

B. The Reaction Products and Analysis

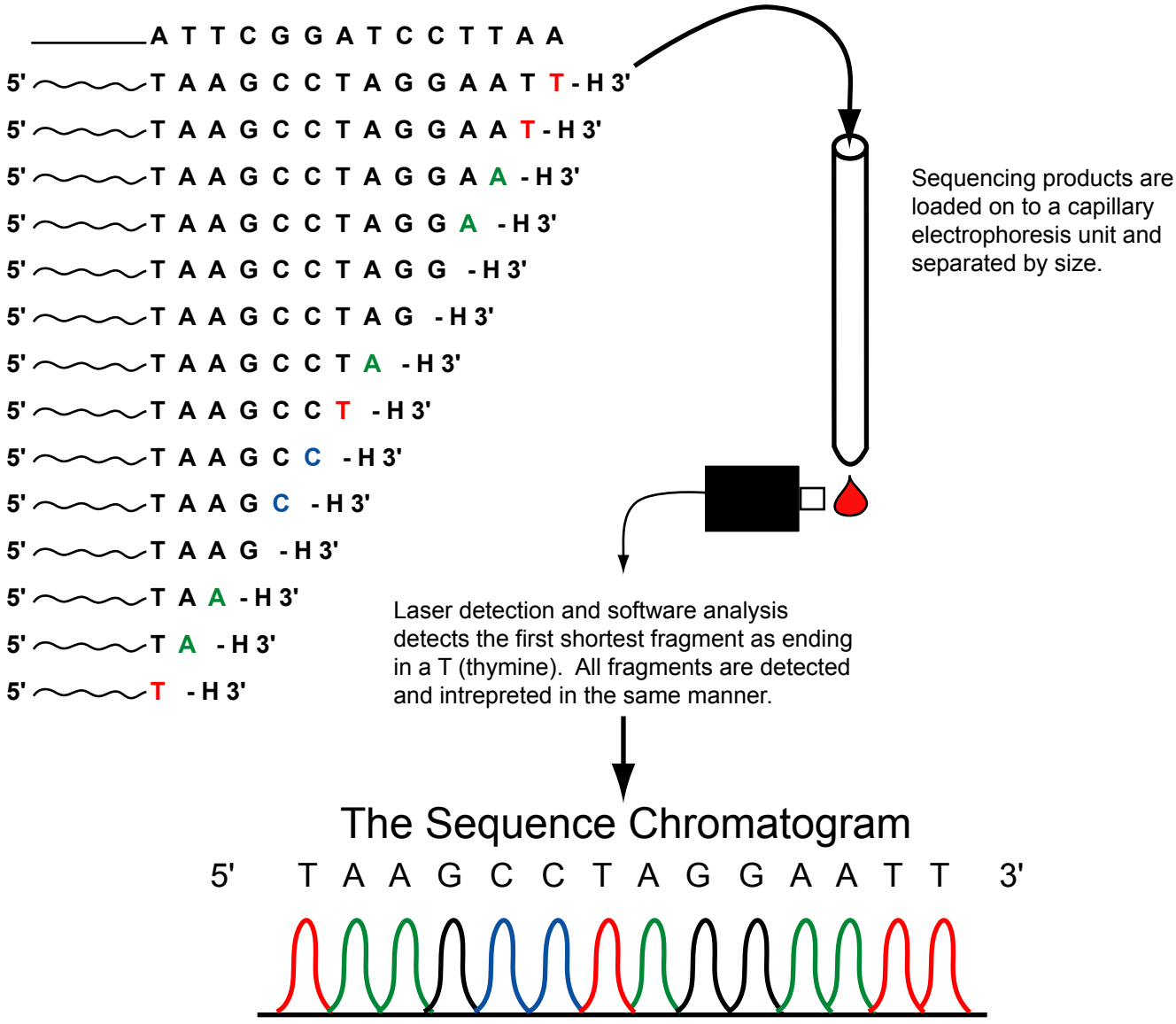
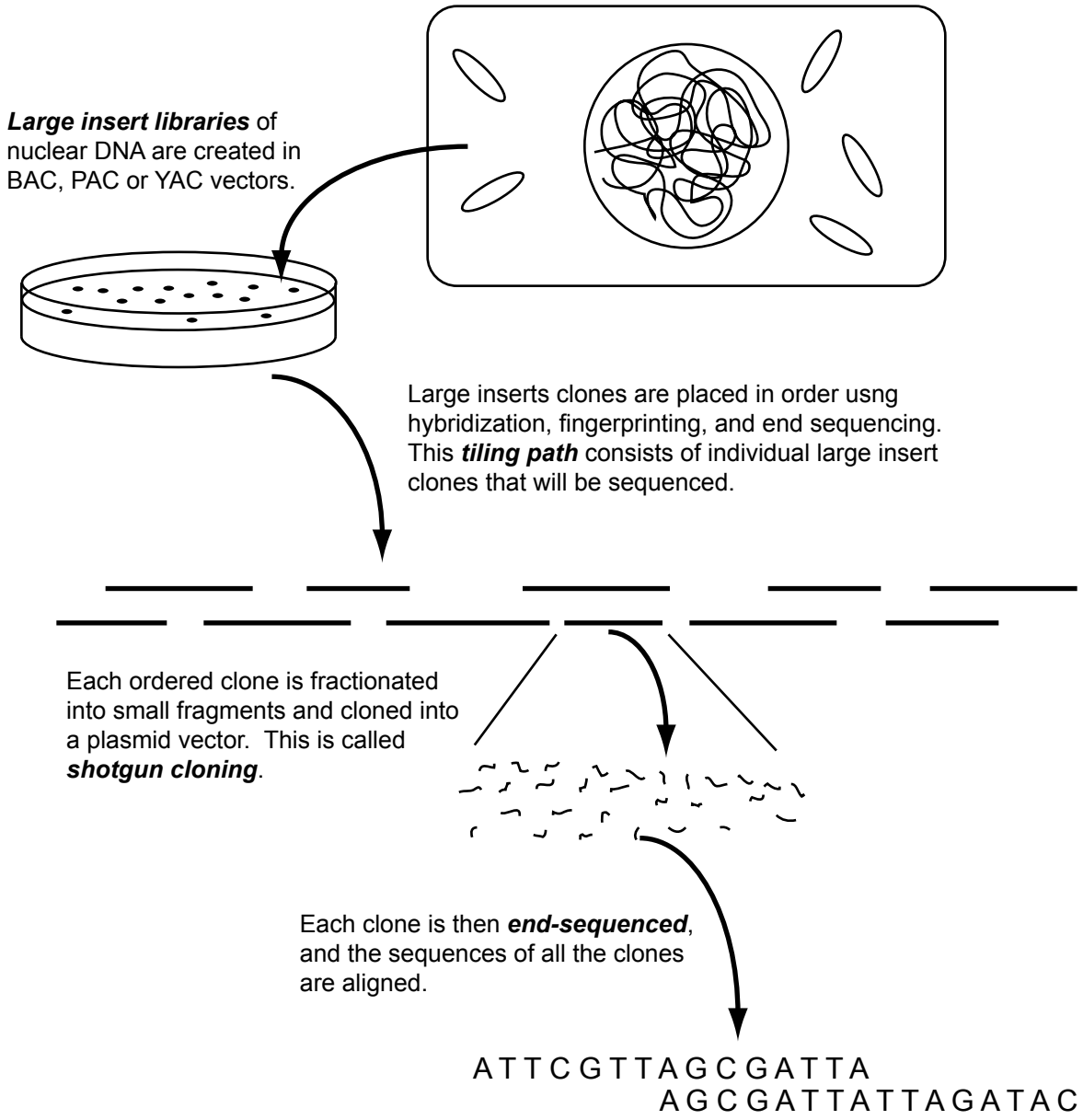


Figure 10. The fluorescent sequencing and laser detectiion process of DNA sequencing.

Hieracrchical Shotgun Sequencing of Genomes

A. The Concept

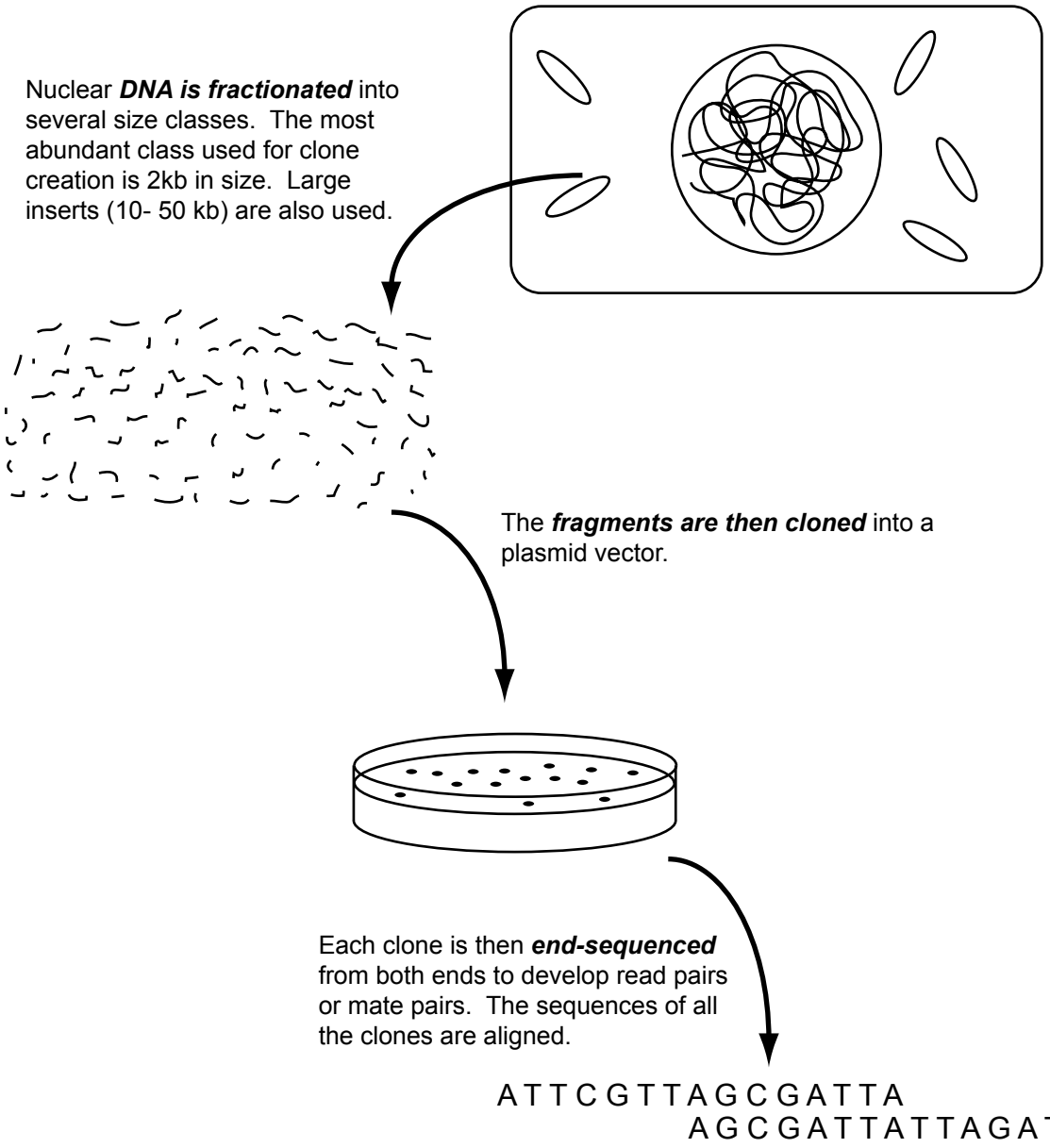
Hierarchical shotgun sequencing requires that large insert libraries be constructed. A series of these clones are ordered by several techniques. Once these clones are ordered, each clone is separately fractionated into small fragments and cloned into plasmid vectors. The plasmid clones are sequenced, and the sequence is assembled. This is the procedure used to sequence the ***Arabidoposis*** genome, and by the public project to sequence the human genome.



Whole Genome Shotgun Sequencing

A. The Concept

Shotgun sequencing requires that random, small insert libraries are created from the total nuclear DNA of the species of interest. A plasmid cloning vector is used for this step. These clones are then sequenced. This step is analogous to the shotgun cloning and sequencing step used for each large-insert clone used in hierarchial shotgun. The sequences of the clones are then aligned. This is the procedure used to sequence the ***Drosophila*** genome, and by Celera to sequence the human genome.



Genome Sequencing

Concept of Genome Sequencing

- Fragment the genomic DNA
- Clone those fragments into a cloning vector
- Isolate many clones
- Sequence each clone

Sequencing Techniques Were Well Established

- Used for the past twenty years
- Helped characterize many different individual genes.
- Previously, the most aggressive efforts
    - Sequenced 40,000 bases around a gene of interest

How is Genomic Sequencing Different???

- The scale of the effort
    - Example
        - Public draft of human genome
            - Hierarchical sequencing
            - Based on 23 billon bases of data
        - Private project (Celera Genomics) draft of human genome
            - Whole genome shotgun sequencing approach
            - Based on 27.2 billon clones
            - 14.8 billion bases

Result:

- Human Genome = 2.91 billion bases

Hierarchical Shotgun Sequencing

- Two major sequencing approaches
  - Hierarchical shotgun sequencing
  - Whole genome shotgun sequencing
- Hierarchical shotgun sequencing
  - Historically
    - First approach
  - Why???
    - Techniques for high-throughput sequencing not developed
    - Sophisticated sequence assembly software not availability
- Concept of the approach
  - Necessary to carefully develop physical map of overlapping clones
    - Clone-based contig (***contig***uous sequence)
  - Assembly of final genomic sequence easier
  - Contig provides fixed sequence reference point
- But
  - Advent of sophisticated software permitted
    - Assembly of a large collection of unordered small, random sequence reads might be possible
  - Lead to **Whole Genome Shotgun** approach

Steps Of Hierarchical Shotgun Sequencing

- Requires large insert library
    - BAC or P1 (bacterial artificial chromosomes)
        - Primary advantages
            - Contained reasonable amounts of DNA
                - about 75-150 kb (100,000 – 200,000) bases
            - Do not undergo rearrangements (like YACs)
            - Could be handled using standard bacterial procedures

Developing The Ordered Array of Clones for Sequencing

- Using a *Molecular Map*
    - DNA markers
    - Aligned in the correct order along a chromosome
    - Genetic terminology
        - Each chromosome is defined as a ***linkage group***
    - Map:
        - Is reference point to begin ordering the clones
        - Provides first look at sequence organization of the genome
- Overlapping the clones
    - Maps not dense enough to provide overlap
    - *Fingerprinting* clones
        - Cut each with a restriction enzyme (*Hin*dIII)
        - Pattern is generally unique for each clone
        - Overlapping clones defined by
            - Partially share fingerprint fragments
    - Overlapping define the ***physical map*** of the genome

Genomic Physical Maps

- Human
    - 29,298 large insert clones sequenced
        - Why so many
            - Genomic sequencing began before physical map developed
            - Physical map was suboptimal
- *Arabidopsis*
    - 1,569 large insert clones defined ten contigs
        - Map completed before the onset of sequencing
        - Smaller genome
            - about 125 megabases

Developing a Minimal Tiling Path

- Definition
    - Fewest clones necessary to obtain complete sequence
- How to find overlaps
- **Fingerprinting**
    - Find share fragments between restriction digested clones
    - Use software to discover overlapping clones

Sequencing Clones of The Minimal Tiling Path

- Steps
    - Physically fractionate clone in small pieces
    - Add restriction-site adaptors and clone DNA
        - Allows insertion into cloning vectors
            - Plasmids current choice
    - Sequence data can be collected from both ends of insert
        - ***Read pairs or mate pairs***
            - Sequence data from both ends of insert DNA
            - Simplifies assembly
            - Sequences are known to reside near each other

Assembly of Hierarchical Shotgun Sequence Data

- Process
    - Data collected
    - Analyzed using computer algorithms
    - Overlaps in data looked for

Confirming the Sequence

- Molecular map data
    - Molecular markers should be in proper location
- Fingerprint data
    - Fragment sizes should readily recognized in sequence data

Whole Genome Shotgun Sequencing (WGS)

- Hierarchical sequencing approach
  - Begins with the physical map
  - Overlapping clones are shotgun cloned and sequenced
- WGS
  - Bypasses the mapping step
- Basic approach
  - Take nuclear DNA
  - Shear the DNA
  - Modify DNA by adding restriction site adaptors
  - Clone into plasmids
    - Plasmids are then directly sequenced
    - Approach requires read-pairs
      - Especially true because of the repetitive nature of complex genomes

WGS

- Proven very successful for nearly all sized genomes
  - Essentially the only approach used to sequence smaller genomes like bacteria
- Early question: ***Is WGS useful for large, complex genomes***?
  - Initially consider a bold suggestion
  - Large public effort dedicated to hierarchical approach
  - *Drosophila*
    - Sequenced using the WGS approach
  - Rice
    - Two different rice genomes sequenced using WGS approach

WGS – Major Challenge 1

- Assembly of repetitive DNA is difficult
  - Retrotransposons (RNA mobile elements)
  - DNA transposons
  - Alu repeats (human)
  - Long and Short Interspersed Repeat (LINE and SINE) elements
  - Microsatellites
- Solution
  - Use sequence data from 2, 10 and 50 kb clones
    - Data from fragments containing different types of sequences can be collected
    - Paired-end reads collected
  - Assembly Process
    - Repeat sequences are initially masked
    - Overlaps of non-repeat sequences detected
    - Contigs overlapped to create supercontigs
  - Software available but is mostly useful to the developers
    - Examples: Celera Assembler, Arcane, Phusion, Atlas

WGS – Major Challenge 2

- For the two sequences approaches
- Assembly is a scale issue
  - WGS approach
    - Gigabytes of sequence data
  - Hierarchical approach
    - Magnitudes less
  - On-going research focuses on developing new algorithms to handle and assembly the huge data sets generated by WGS

Pyrosequencing in Picolitre Reactors

Pyrosequencing reagents

- DNA template (DNAn)
- DNA polymerase
- A dideoxynucleotide
  - dNTP
  - deoxyadenosine thio triphospate substitutes for dATP
- ATP sulfurlyase
- Adenosine 5’ phosphosulfate (APS)
- Luciferase
- Luciferin
- Apyrase

Pyrosequencing reactions

(1) $\text{DNAn} + \text{a dNTP} \xrightarrow{\text{DNA Polymerase}} \text{DNAn+1} + \text{PPi}$

(2) $\text{PPi} + \text{APS} \xrightarrow{\text{ATP Sulfurylase}} \text{ATP}$

(3) $\text{ATP} + \text{Luciferin} \xrightarrow{\text{Luciferase}} \text{Oxyluciferin} + \text{Light}$

(4) $\text{dNTP} \xrightarrow{\text{Apyrase}} \text{dNDP} + \text{dNMP} + \text{phosphate}$

(5) $\text{ATP} \xrightarrow{\text{Apyrase}} \text{dADP} + \text{dAMP} + \text{phosphate}$

- Important points about pyrosequencing reaction
    - One nucleotide is introduced at a single time
    - Data is collected from a charge coupled devise (CCD)
        - Used to detect light emission
        - A single photon of light is detected for each nucleotide introduced
    - After reaction is complete
        - New set of reagents (different nucleotide) is introduced
        - Repeated many steps to collect sequence data

454 Life Sciences DNA Sequencing System

- Utilizes pyrosequencing to collect sequence data

Preparation of sequencing template

- Genomic DNA is
    - Sheared
    - Adapators added to the end
    - Made single-stranded
    - SS DNA is bound to a bead
    - Single bead/DNA combination encapsulated in emulsion
    - DNA is duplicated on the bead in the emulsion
    - Each emulsion DNA bead is a single DNA reaction vessel

Fiberoptic plates

- Plate contains 1.6 million wells
- CCD mounted to back of plate
- One DNA/bead emulsion is loaded per well

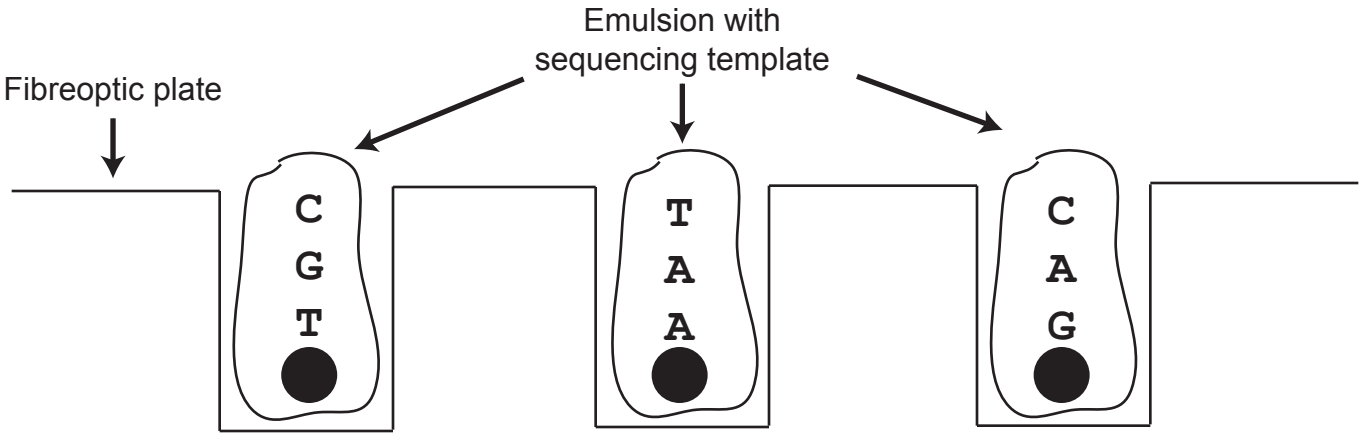
How it works

- Pyrosequencing reagents (one dNTP at a time) are added
- CCD collects sequence results for each well for that dNTP
- Residual sequencing reagents washed out
- New reagent for second dNTP added
- Process continues until finished

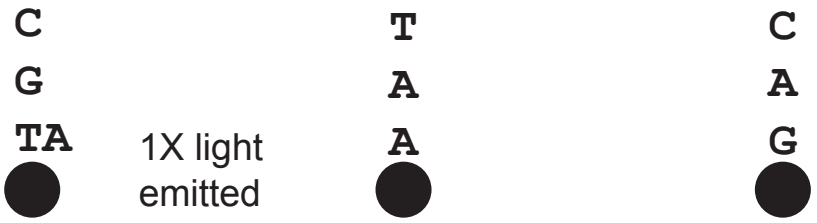
Major constraint

- Read length
    - ~400 bases
        - But overcome by better assembly process

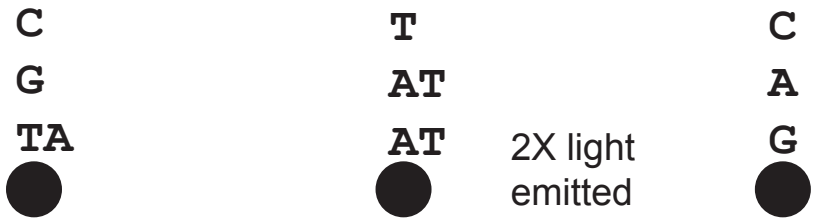
454 Sequencing Reaction Principles



Step 1: add dATP
and other reagents



Step 2: add dtTP
and other reagents



Step 3: add dCTP
and other reagents

Roche 454 Sequencing

1. Make the sequencing bead

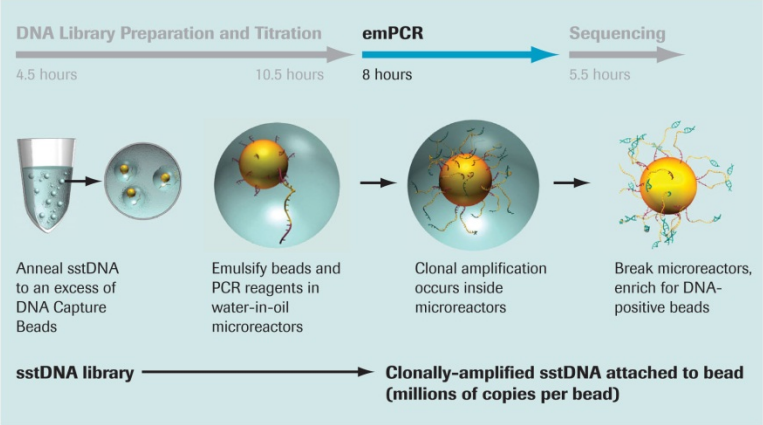


Figure 4: Overview of emulsion-based clonal amplification (emPCR) with the Genome Sequencer 20 System.

2. Insert into sequencing well

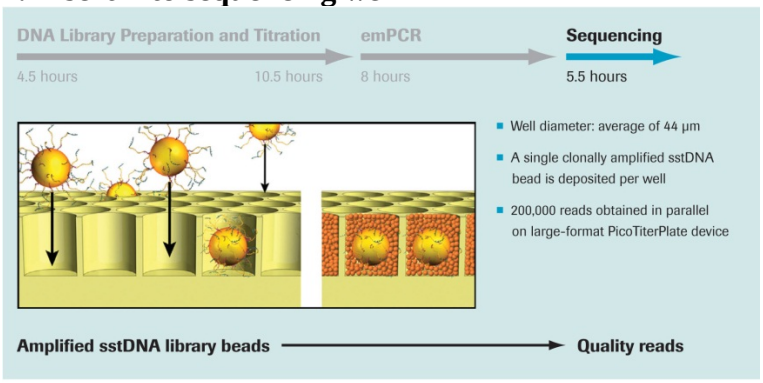
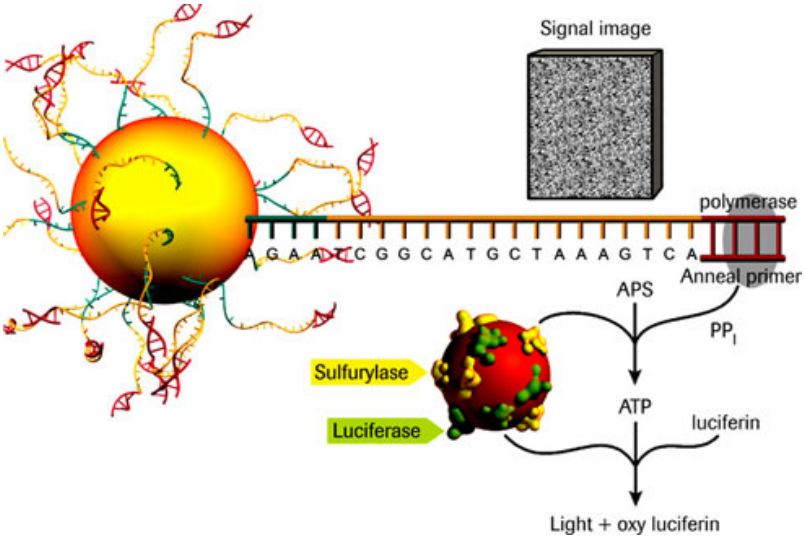


Figure 5: Deposition of DNA beads into the PicoTiterPlate device.

3. Perform sequencing reaction



Sequencing by Synthesis - Illumina System (now owned by Illumina)

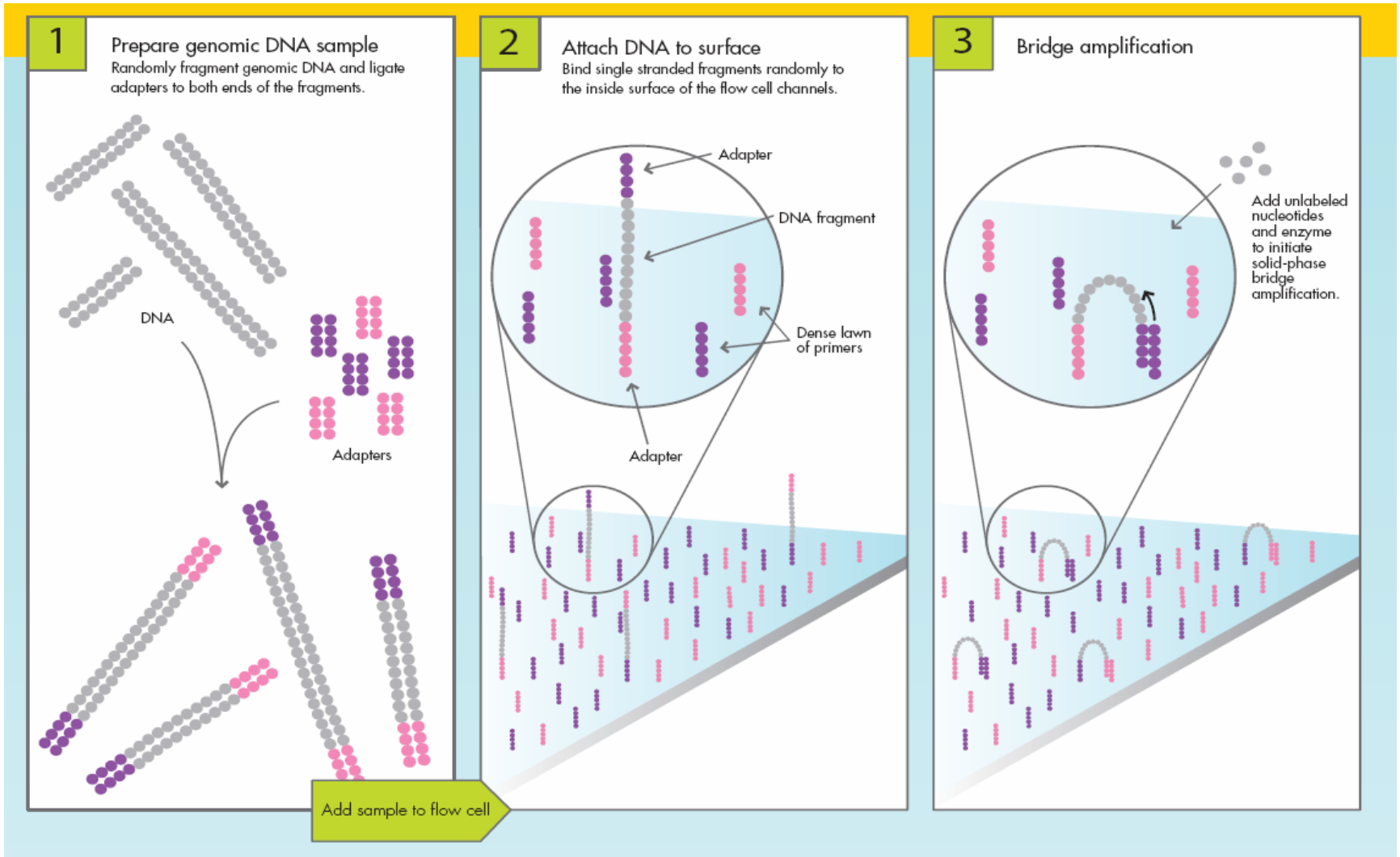
Basic Steps

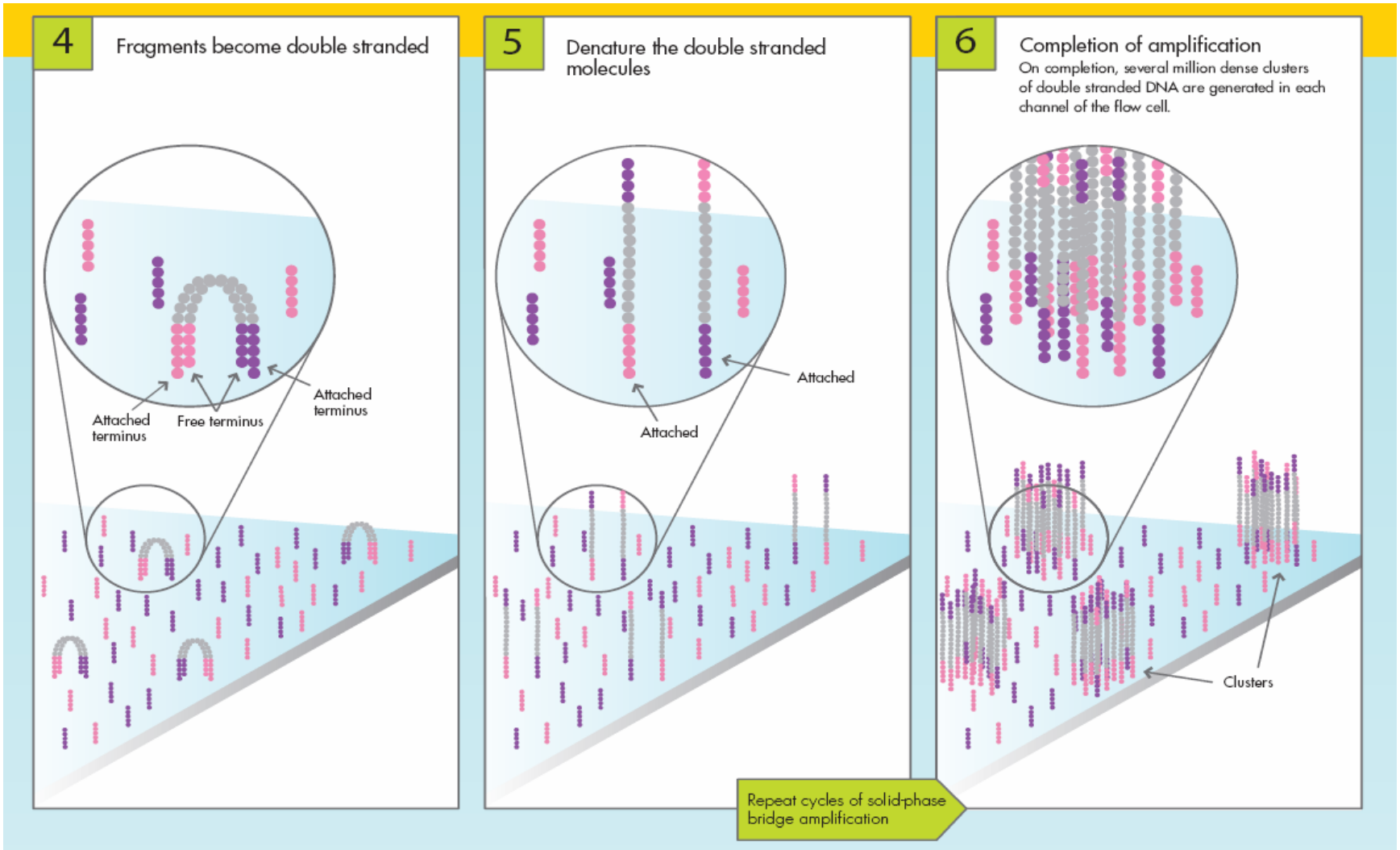
- Sequencing matrix contains many copies of two different primers
- Sequencing targets are created in clusters
- Sequential addition/detection of fluorescent labeled nucleotide from each target cluster

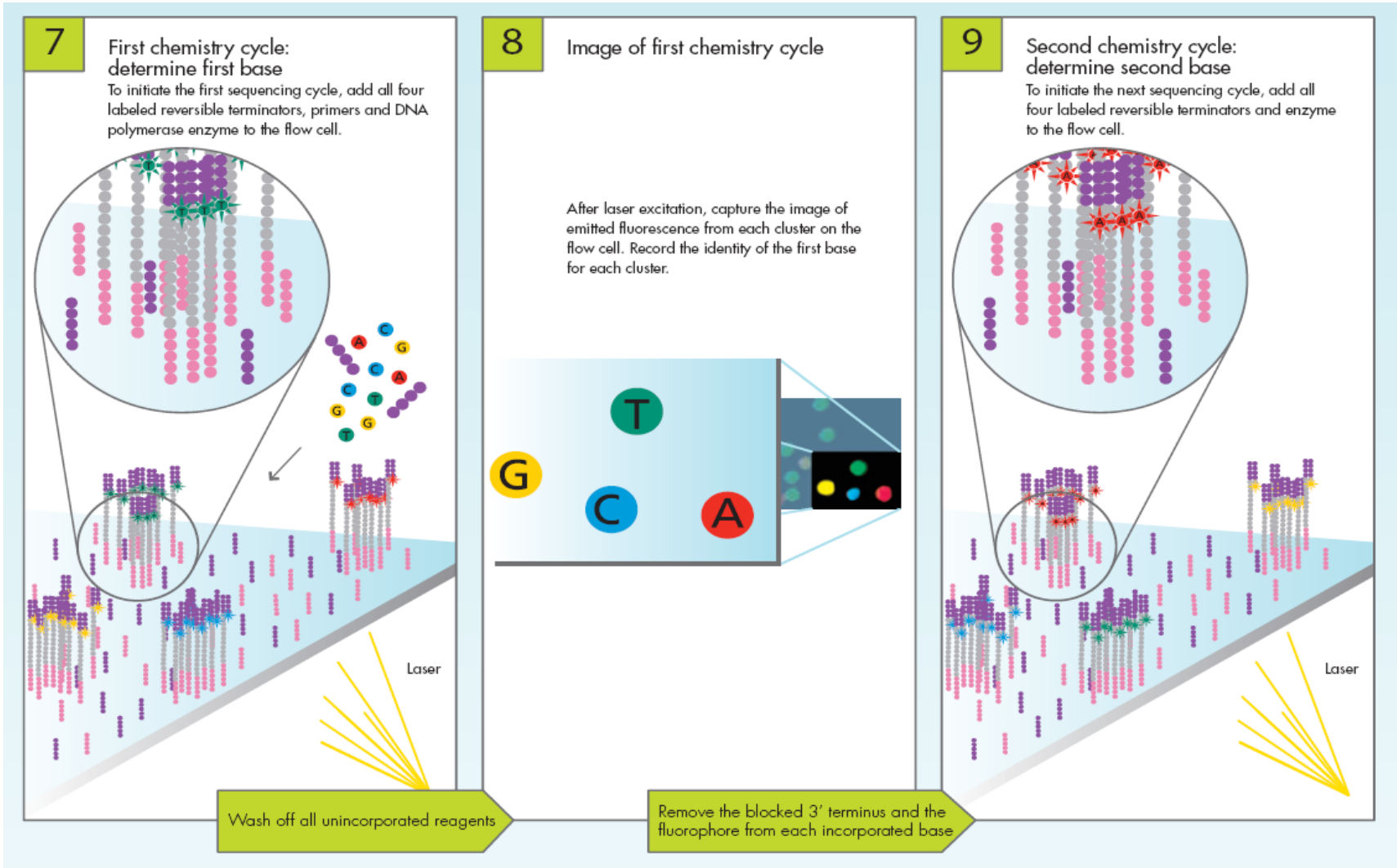
Detailed Steps

- **Preparing DNA**
- Sheared DNA is prepared
- Adaptors homologous to the two primers are attached to ends
- **Building Clusters**
    - Adaptor/target DNA is made single stranded and bound to matrix
    - Adaptor/target DNA is bridged to bound primer by hydrogen bonding
    - Solid phase bridge amplification creates double stranded product
    - Double stranded product is made single stranded and now two single strands are attached nearby on matrix
    - Solid phase bridge amplification cycle is repeated to create local cluster of identical sequencing templates
- **Sequencing**
    - Chemistry cycle begins by addition of four labeled ***reversible*** terminators and DNA polymerase
    - Laser detection records the base at each cluster
    - 25-30 chemical cycles are run
    - Sequence data is collected at each cluster site

Solexa Sequencing by Synthesis Technology

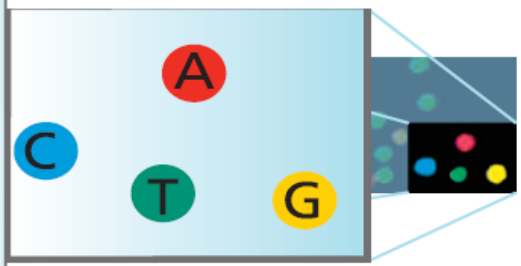



4
Fragments become double stranded
Attached terminus
Free terminus
Attached terminus
5
Denature the double stranded molecules
Attached
Attached
6
Completion of amplification
On completion, several million dense clusters of double stranded DNA are generated in each channel of the flow cell.
Clusters
Repeat cycles of solid-phase bridge amplification


7
First chemistry cycle: determine first base
To initiate the first sequencing cycle, add all four labeled reversible terminators, primers and DNA polymerase enzyme to the flow cell.
Laser
Wash off all unincorporated reagents
8
Image of first chemistry cycle
After laser excitation, capture the image of emitted fluorescence from each cluster on the flow cell. Record the identity of the first base for each cluster.
T
G
C
A
Remove the blocked 3' terminus and the fluorophore from each incorporated base
9
Second chemistry cycle: determine second base
To initiate the next sequencing cycle, add all four labeled reversible terminators and enzyme to the flow cell.
Laser

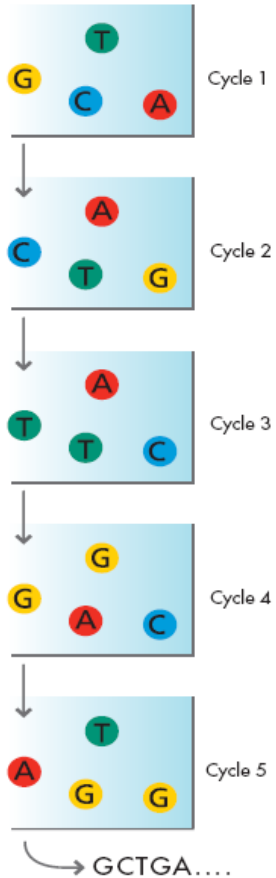
10 Image of second chemistry cycle is captured by the instrument

After laser excitation, collect the image data as before. Record the identity of the second base for each cluster.

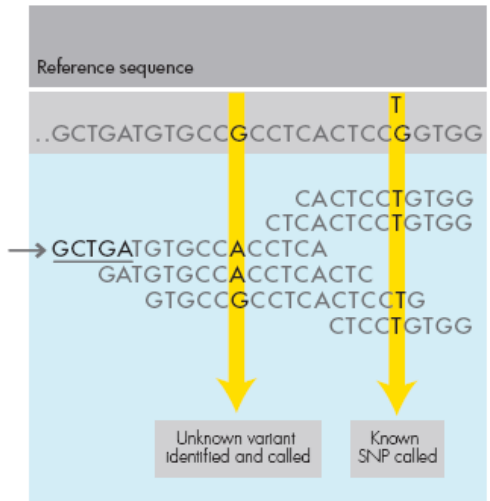


11 Sequence read over multiple chemistry cycles

Repeat cycles of sequencing to determine the sequence of bases in a given fragment a single base at a time.



12 Align the new data to a reference and identify sequence differences



Illumina DNA Sequencing Throughput (from company literature)

Genome Analyzer – (standard equipment until March 2011)

Read length	Run time (days)	# of reads/flow cell (million)	High-quality output (Gbases)	Base calls with quality > q30
1 x 35 bp	2.5	138-168	4.5-6	70-85%
2 x 35 bp	5	138-168	9.5-11.5	70-85%
2 x 50 bp	6.5	138-168	13.5-16.5	70-85%
2 x 75 bp	9.5	138-168	20.5-25	>70%

HiSeq 2000 Dual Flow Cell

Read length	Run time (days)	# of reads/flow cell	High-quality output (Gbases)	Base calls with quality > q30
1 x 35 bp	2	4 billion (max)	95-105	na
2 x 35 bp	5.5	6 billion (max)	270-300	85%
2 x 100 bp	11	6 billion (max)	540-600	80%

Single Polymerase Real Time DNA Sequencing

Developed by Pacific Biosciences

- Native rate of DNA replication
  - ***1000 nucleotides/second***
- Pacific Biosciences system
  - Sequences occurs at the rate of ***10 nt per second***
    - ***MUCH FASTER THAN ALL OTHER SYSTEMS***

Principle

Reaction Cell

- A single DNA polymerase is immobilized on the bottom of a reaction cell
- Φ29 DNA polymerase is used
- Each sequencing plate contains ~100,000 individual cells
  - Each holds only a single DNA molecule

Chemistry

- A phospholinked dNTP is used
  - Each dNTP contains a different fluorophore
- During sequence
  - A single labeled dNTP enters the polymerase
  - dNTP held in place shortly
  - Fluorescence signal is emitted in the ZMW for a short period of time
  - dNTP leaves and new dNTP enters

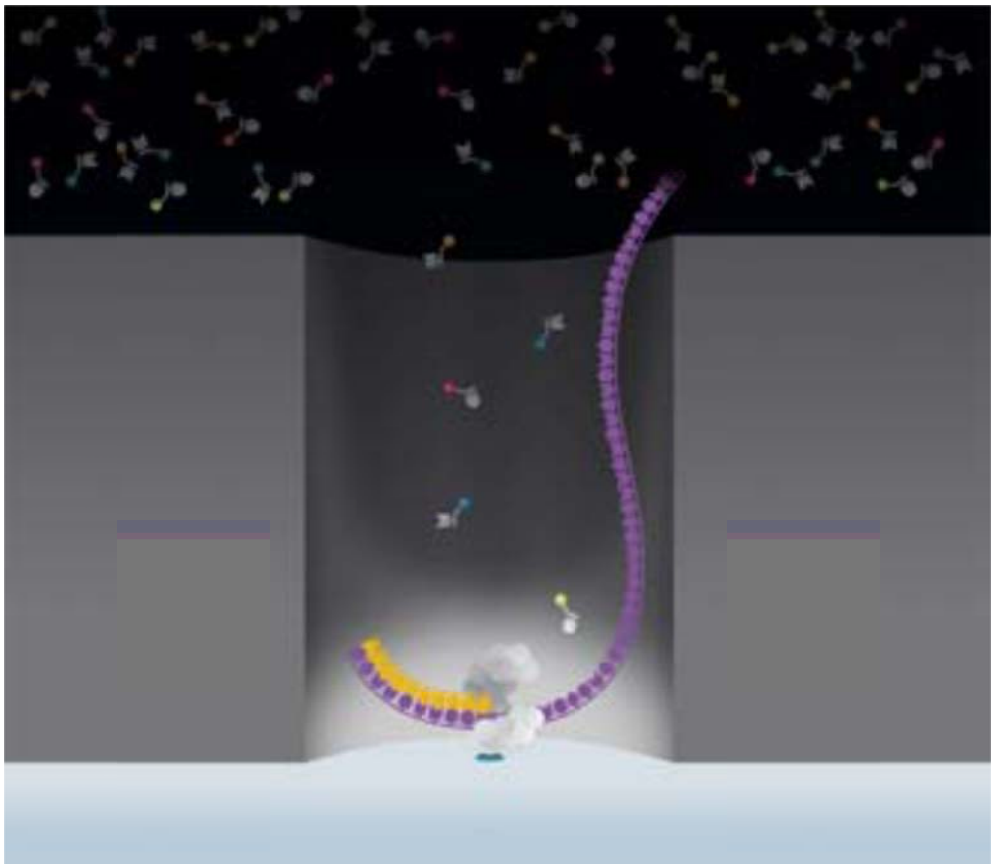
Detection and sequence determination

- Fluorescence signals for each ZMW collected
    - Data is collected as a movie of the sequential signals
        - Each individual signal is measured as a short pulse of light
    - Successive fluorescence signal data is collected
    - DNA sequence of single molecule is determined by sequence of light pulses

Images and Notes Below From:

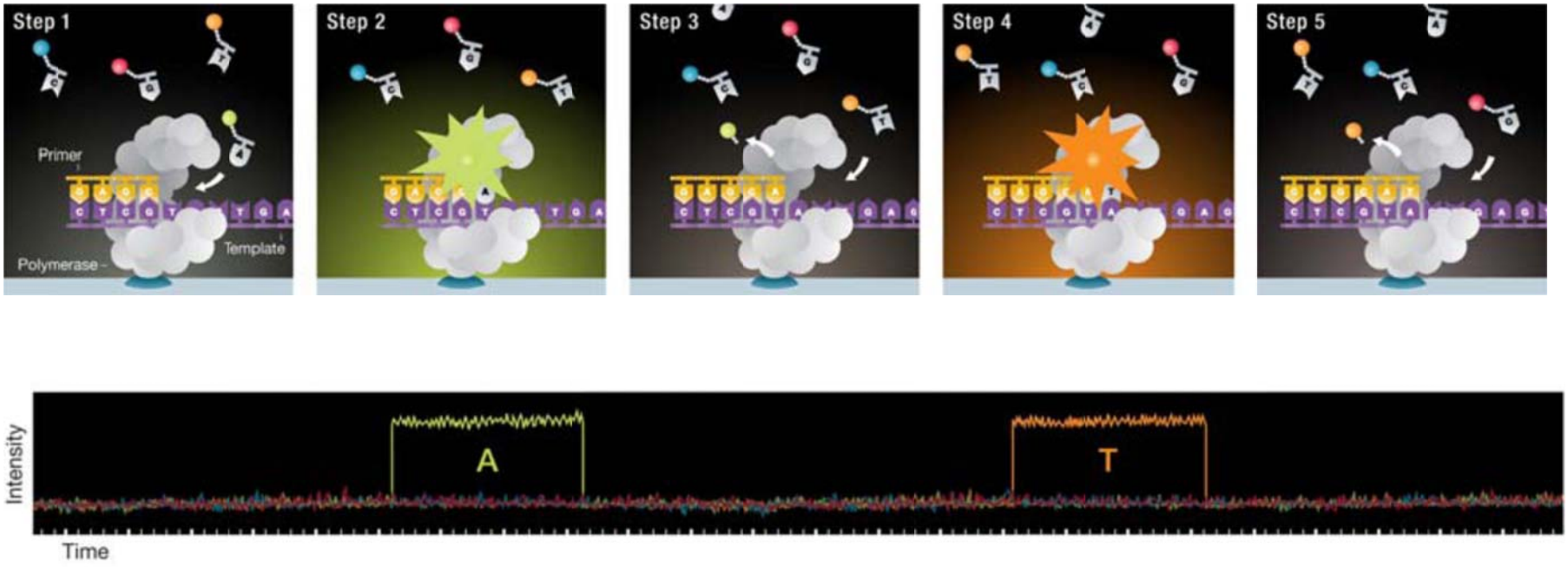
Pacific Biosciences Technology Backgrounder (11/24/2008)

Title: Pacific Biosciences Develops Transformative DNA Sequencing Technology: Single Molecule Real Time (SMRT) DNA Sequencing



ZMW (Zero-mode waveguide) with Φ29 DNA polymerase and DNA template

Single Polymerase DNA Sequencing



Step 1: Fluorescent phospholinked labeled nucleotides are introduced into the ZMW.
Step 2: The base being incorporated is held in the detection volume for tens of milliseconds, producing a bright flash of light.
Step 3: The phosphate chain is cleaved, releasing the attached dye molecule.
Step 4-5: The process repeats.

Potential Advantages

Speed

- 10 nt per second

Length

- 1000-2000 nt
- This is a claim
    - Not fully proven

Assembly

- Much easier to assembly longer fragments

Cost

- Company claim
    - $10/human genome

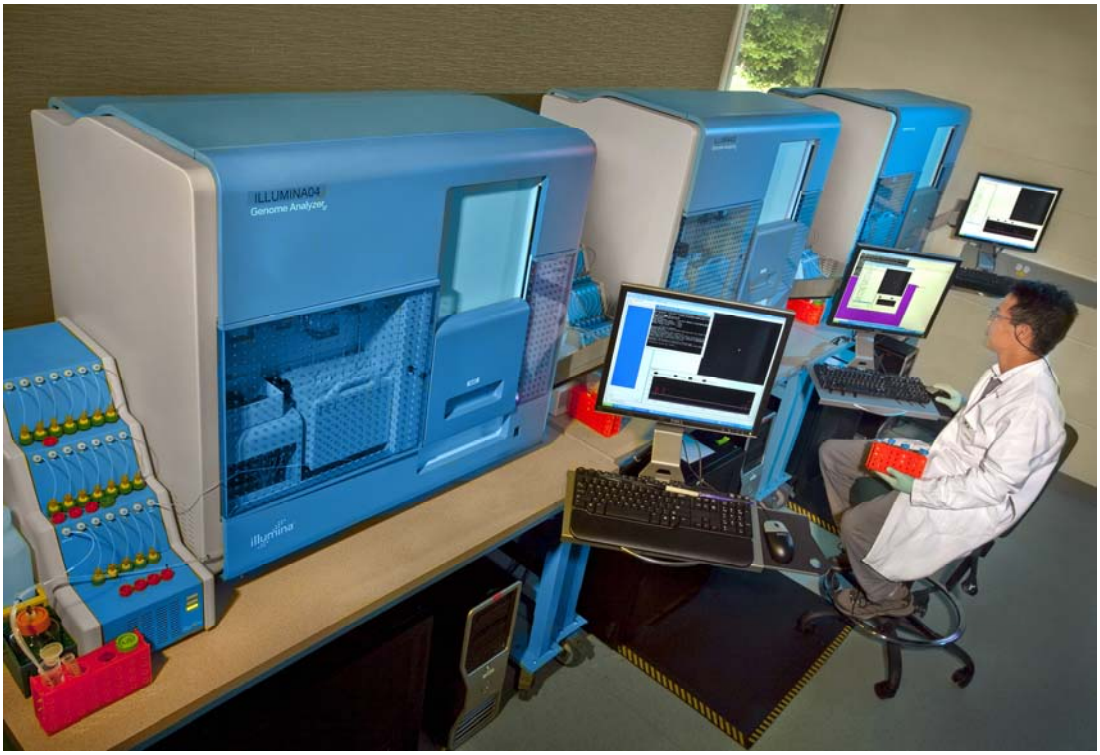
Department of Energy
Joint Genome Institute

Sequencing Productivity

	Total bases (billions) by platform			
Year	**Sanger**	**454**	**Illumina**	**Operating hours**
2008	20.3	41.2	64.0	7,704
2009	20.6	170.6	812.7	8,626
2010	4.1	360.6	5,676.1	8,712
2011 (3 quarters)	Retired	122.7	16,004.0	6,552
Total	46.0	695.1	22,556.8	

THEN: DOE Sanger Sequencing Equipment Room

NOW: DOE Illumina Sequencing Equipment Room



Sequencing the Gene Space: An alternative to whole genome sequencing

Background

Major goal of genome sequencing

- Define the gene set
- Large genomes create a problem
    - High ratio of non-coding to coding DNA
        - ex. Human genome is 3% coding (or gene-based) DNA

Not all genomes are equally valued

- Many crop species have little support for complete genome sequencing
    - Why???
        - Not model species like Arabidopsis, rice or *Medicago truncatula* (legume model)
        - Genomes are complex with large amount of repetitive sequences

What Nucleic Acids Are Sequenced???

DNA - the most common nucleic acid sequenced.

- Isolate total genomic DNA
- Create some sequencing library
- Sequenced using the Sanger approach or a NexGen

Other subclasses nucleic acids are also sequenced

- Studies just a fraction of the genome
    - Two alternate nucleic acid pools
        - Exons
        - mRNA from specific tissues

Exome sequencing

- Ex**ome** - all of the exons of a genome
    - Ex**omics** – the study of all of the exons of a genome at the same time

Why focus on the exons ?

- Most mutant phenotypes are the result of mutations in exons.
    - Important mutations can be discovered and studied
    - More efficient if your goal is to look for mutations in the coding region
    - Human genome
        - Only 3% of the genome is composed of genes
        - On average only 1/3 of the gene consists of exons
            - 1% of human genomic DNA is exons

Exon capture - an approach to collect exon DNA

- Requires a reference genome sequence with gene models
  - Gene modeling defines the exon and intron boundaries for each gene
- Long oligonucleotides highly similar to a part of a specific human exon are utilized to capture just exon sequences

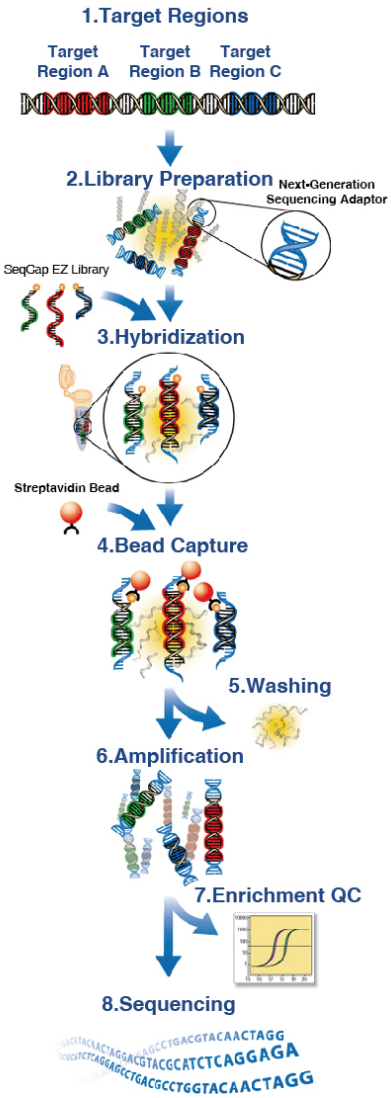
NimbleGen – developer of exon capture technology.

- NimbleGen SeqCap EZ Exome V2.0 solution system
  - 2.1 million oligonucleotides
  - Target ~300,000 exons
  - ~30,000 genes
  - Average - 10 oligonucelotides (=oligos) per exon
  - Captures ~36.5 Mb of human genomic DNA
- NimbleGen Sequence Capture 2.1M Human Exome Array
  - ~2.1 million long oligos
  - Targets 180,000 exons and 551 micro RNA exons
  - Captures ~50 Mb of human genomic DNA

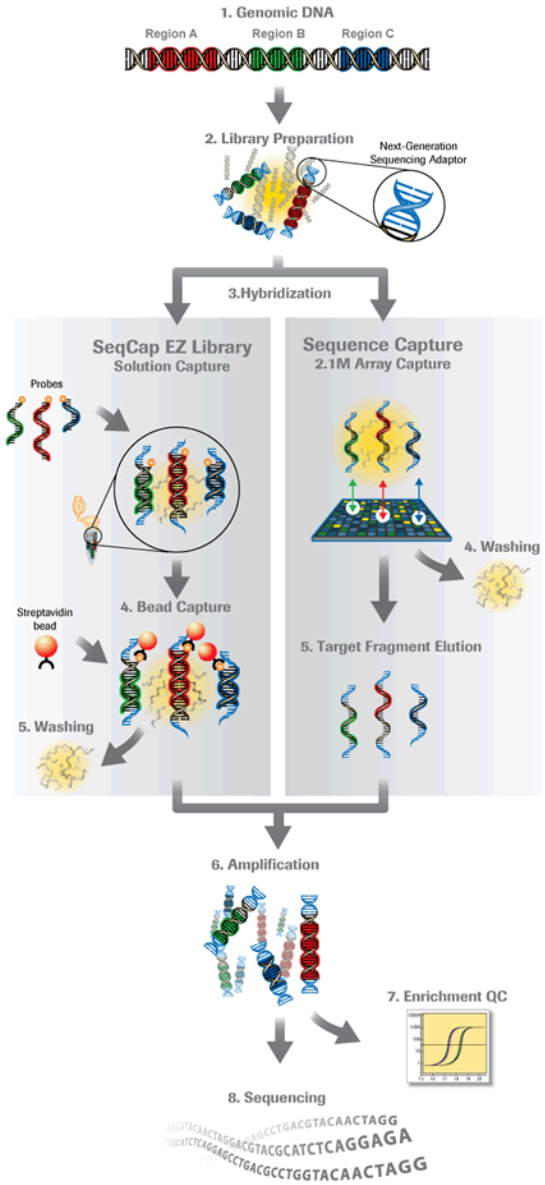
How does exome capture work?

- Fractionated into small fragments
- Denature fragments (made them single-stranded)
- Hyrbridize fragments to oligos in solution or on array
  - Fragments complementary to the oligos are bound to the DNA.
- DNA bound to the oligos is recovered
- DNA sequenced
  - Data analyzed for mutant discovery

Solution Exome Capture



Array Exome Capture



Sequencing RNA

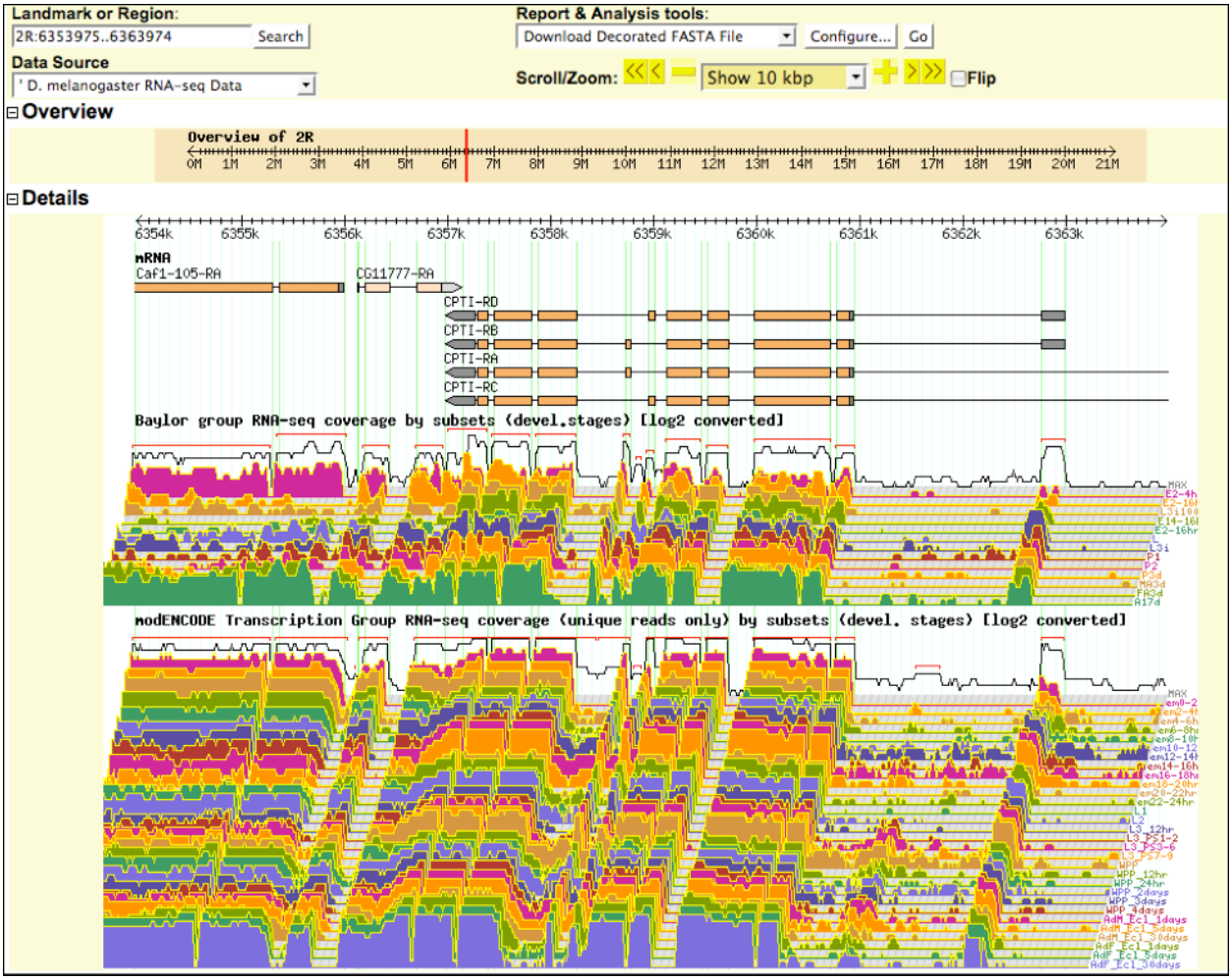
- RNA fraction studied since early days of molecular genetics
    - Why?
        - All expressed genes found in the RNA fraction.
- The first method used to consider the genes expressed at a specific stages was EST sequencing
    - Required that development of a cDNA library (DNA copies of mRNA sequences)
    - Many cDNA clones sequenced using the Sanger sequencing technique.
        - Data = EST, ***expressed sequence tags***
- Challenge - capture all of the mRNA in a specific tissue.
    - Never achieved
        - Abundantly expressed mRNAs were predominant
        - Rare mRNAs underrepresented

RNA-seq (or RNA sequencing)

- Limitation overcome by using next generation sequencing
- Massive amounts of sequence data ensures that all of the mRNA transcripts will be sequenced
    - Copy number of a sequence found in sequence collection is proportional to the number of expressed copies of that specific gene.
    - The expression pattern of specific genes can now be evaluated in detail

RNA-seq results

Notice only exon sequences are represented in the RNA-seq output



How much sequence do you need for a research project?

- Important question to address this research question
  - Low coverage (limited amount of sequence data) may be enough
    - But NexGen sequencing produces massive amounts of data

Bar coding and pooled sequencing

- A method to leverage the large output of NexGen sequencing
  - Cost of NexGen sequencing is spread over many samples
- How is it done?
  - Unique sequence tags added to fragments during library preparation
  - Multiple libraries are pooled and sequenced in a single lane
  - Sequences containing the same tag are evaluated together